### Title
Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis




### Authors
Samberg, Rachael Gayza
Hennesy, Cody



### Publication Date
2019-10-15

Peer reviewed

CHAPTER 17

# Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis

*Rachael G. Samberg*
*Cody Hennesy*[1]

Imagine you are working with two digital humanities scholars studying post-WWII poetry, both of whom are utilizing a single group of copyright-protected works. The first scholar has collected dozens of these poems to closely analyze artistic approach within a literary framework. The second has built a personal database of the poems to apply automated techniques and statistical methods to identify patterns in the poems' syntax. This latter methodology—in which previously unknown patterns, trends, or relationships are extracted from a collection of textual documents—is an example of "computational text analysis" (CTA),[2] also commonly referred to as "text mining" or "text data mining."[3]

In accessing, building, and then working with these collections of texts (or "corpora" to use the jargon of the digital humanities), both scholars are exercising rights and making elections that carry legal impact. Indeed, they may not even be aware of the choices they can or must make:

- From a copyright fair use perspective, does it matter whether a scholar compiles poems to read (or "consume") or, like the CTA scholar above, uses algorithms to mine information within them (often referred to as "non-consumptive" analysis)?
- How does an added layer of university database licensing, a publisher-provided API (application programming interface), a university archives agreement, or a website's "terms of use" fit into a CTA researcher's protocol for content access, collection, and analysis? When might conditions of those agreements or tools bear upon the researchers' fair use rights?
- And what should researchers know about whether they can subsequently share the corpus they use or create or republish excerpts from it in their scholarship?

Guiding scholars in addressing these issues before they build their research corpora can help them avoid unexpected pitfalls, particularly when a CTA scholar must grapple with unique copyright scenarios. Currently, many CTA researchers programmatically access and download copyright-protected works—even when it potentially violates copyright, licenses, privacy, or computer fraud law—because it is technically feasible. Few of these researchers are malicious in intent; rather, they may lack the necessary training or support to safely navigate the obscure regulatory environment of the field.

Already, some guidance on the legal issues arising within CTA has been created for European Union researchers.[4] Resources offering similar assistance under a US legal framework are just beginning to emerge.[5] This chapter attempts to build upon such input in an effort to address CTA support from a researcher's perspective. Here, we survey copyright and other legal terrain affecting CTA, exploring where these legal issues intersect with CTA methodologies to illuminate pain points for researchers. We then sketch a scholarly workflow that unites law and CTA practice—a roadmap meant to be both adoptable and adaptable by scholars in the field.

# Framing the Issues

## *Copyright, Fair Use, and Computational Text Analysis*

Modern researchers are often copyright savvy and understand that authors (including themselves) have protectable rights. Not all researchers, however, are familiar with what the Constitution intended copyright to *encourage*—the progress of science and useful arts.[6] Indeed, as the Second Circuit Court of Appeals recently observed, "While authors are undoubtedly important intended beneficiaries of copyright, the ultimate, primary intended beneficiary is the public, whose access to knowledge copyright seeks to advance by providing rewards for authorship."[7]

In implementing the Constitution's directives, Congress therefore built exceptions into the Copyright Act. Congress created these exceptions to achieve the aims of advancing public knowledge and understanding by limiting the scope of copyright holders' exclusive rights. One of the strongest such limitations is the right of fair use, codified in 17 U.S.C. § 107. It provides that the fair use of a copyrighted work "for purposes such as criticism, comment, news reporting, teaching,…scholarship, or research is not an infringement."[8] In Section 107, Congress offered four non-exclusive factors to consider in making a fair use determination:

1. the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
2. the nature of the copyrighted work (with use of factual works more likely to be fair under this factor than use of fictional works that come closer to the "core of creative expression");[9]
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
4. the effect of the use upon the potential market for or value of the copyrighted work.

Evaluating whether a given use of copyrighted material is "fair" requires balancing these four factors on a case-by-case basis.[10]

As a practical matter, courts often tend to give particular weight to factors one and four, which are, themselves, interconnected, given that as the

character of a new use becomes more distinct from the original, the less impact the new use would have on the market for that original.[11] Because factors one and four have a special importance, particularly in the adjudication of research-related uses, it is important to dig a bit deeper into how they pan out:

- Factor one's consideration of the "character" of the use prompts courts to inquire whether the new use "merely supersedes the objects of the original creation, or instead adds something new, with a further purpose or different character."[12] Significantly, use of a copyrighted work need not modify or augment the original to be transformative.[13] Rather, the use need only be productive and employ the material in a different manner or for a different purpose from the original—adding "new information, new aesthetics, new insights, and understandings."[14] The more transformative the new work, the less significant countervailing aspects (like commercialism) would be under this factor.[15]
- Factor four requires courts to determine whether the new use of a work would "materially impair the marketability" of the original and whether the new form "would act as a market substitute" for the original.[16] It is significant to note that the focus is not on whether the secondary use suppresses or eliminates the market for the original or its potential derivatives but rather "whether the secondary use *usurps* the market of the original work."[17] In other words, a mere adverse market effect alone is not enough for a fourth factor to weigh against an overall finding of fair use. This is critical for new scholarly works like criticism or parody because merely suppressing demand for the work being criticized does not overcome a fair use determination.[18]

While CTA researchers may be familiar with some of these contours of fair use, exactly how fair use relates to their specific computational methods or their plans to publish can be quite complex. CTA allows users to, among other things, "discern fluctuations of interest in a particular subject over time and space by showing increases and decreases in the frequency of reference and usage in different periods and different linguistic regions."[19] Yet, to achieve a sufficient corpus for reliable and thorough analysis, scholars must often create intermediate downloads of materials from various sources *en masse*—not to read them, but to perform compu-

tations on them. Is creating a corpus or database of copyrighted materials for CTA fair use?

Several courts have considered the intersection of full text searching a corpus and fair use, each finding non-consumptive text mining to be fair.[20] Particularly instructive are the Court of Appeals' holdings in *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) (*HathiTrust 2014*): Scanning and creating a database of digitized materials so that users could conduct full text searching within the content, rather than read that content, was both transformative and a fair use overall. In *HathiTrust 2014*, a collection of authors and authors associations sued HathiTrust,[21] certain of its member universities and university presidents for copyright infringement. The basis of their claims was the fact that, pursuant to a relationship with Google, HathiTrust received digital copies of nearly ten million books—the majority of which were still in-copyright. HathiTrust then made these books available for full-text searching (and, inherently, CTA) essentially within a "black box"—i.e., without the researcher being able to read or "consume" the book. For instance, HathiTrust permitted users of the HathiTrust Digital Library (HDL) to search HDL to determine where in a book (i.e., on which page numbers) and how often a search term appeared but without a user window into the text.[22] As the court noted, "HDL does not display text from the underlying copyrighted work (either in "snippet" form or otherwise). Consequently, the user is not able to view either the page on which the term appears or any other portion of the book."[23]

The court found this arrangement to be fair use, notably because the textual analysis that HDL enabled was transformative under the first fair use factor. The court explained:

> An important focus of the first factor is whether the use is "transformative." A use is transformative if it does something more than repackage or republish the original copyrighted work. The inquiry is whether the work "adds something new, with a further purpose or different character, altering the first with new expression, meaning or message…." [citations omitted].[24]

In turn, under this standard:

> The creation of a full-text searchable database is a quintessentially transformative use…. [T]he result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn. Indeed, we can discern little or no resemblance between the original text and the results of the HDL full-text search.[25]

In fact, full-text searching was considered so transformative that the first factor outweighed any of the other three factors that might have otherwise leaned against fair use.[26] Still, the court observed that other factors also supported a determination of fairness: copying of books in their entirety (factor three) was necessary to enable full-text search functionality and reliability, and full-text searching was not a market substitute for purchasing and reading the original books (factor four).[27]

The Second Circuit Court of Appeals also ruled in *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015)[28] (*Google Books 2015*) that Google Books' creation of a full-text searchable database and "Ngram Viewer" (discussed below) were fair uses. In addition, allowing users to view three-line snippets of the underlying works, to provide context for where desired phrases appear, was similarly fair use. As Jockers, Sag, and Schultz had described before *HathiTrust 2014 and Google Books 2015* were decided, "Scanning words from library books to make a search index, or to compile a list of word frequencies, does not interfere with the rights of the author. These uses simply convert masses of text into metadata."[29] Indeed, the cases that followed affirmed both the transformativeness of this arrangement and the ability of digital libraries to leverage fair use in the creation of CTA interfaces and mechanisms.

## Implications of Case Law, and Limits of a Black Box

New tools that rely on transformativeness and fair use continue to be developed. Increasingly, there are options for users to access derived downloadable datasets representing texts (e.g., ngrams), or to use web tools to visualize trends from digital collections without exposing the underlying texts. Further, the HathiTrust Research Center (HTRC) provides secure computing environments, or "data capsules," where researchers can work with public domain texts (with plans to expand access in the future to in-

clude in-copyright works) from the HDL on a virtual computer, without allowing for the release of full-text data.[30]

Yet, these tools also introduce digital literacy challenges because they transform works in ways that an aspiring CTA researcher may find puzzling. For the researcher hoping to bulk download .txt or PDF files of a particular corpus, for example, it may be confusing to encounter data from *JSTOR Data for Research*[31] or the *HTRC Extracted Features Dataset*,[32] where the words from articles and books are not available in their original order, but rather as word-counts, ngrams, tokens, or features. Learning more about both the copyright restrictions governing access to the original documents and the fair use considerations that have enabled these new forms of access can help users determine whether these pre-compiled corpora suit their research purposes.

For some research projects, however, word-counts, ngrams, or tokens may be insufficient. One major disadvantage for researchers who download Google Books Ngrams,[33] for example, is that they will be unable to clearly define or articulate the boundaries of their corpus[34]—that is, to identify which books are and are not included or to ask questions of specific volumes.[35] Further, there are a variety of CTA methods that require words to be available in their original order, and that would, therefore, be inappropriate to attempt using derived downloadable datasets or other "bag of words" models.[36] At the simplest level, for example, one might wish to track the occurrence of an exact eight-word phrase in a corpus, a technique that would be impossible if one is unable to access any more than three words in a row (trigrams) from copyrighted texts. Since derivative datasets introduce design artifacts that can be difficult to fully understand or explain, many researchers prefer a corpus that closely resembles the original, readable collection of texts.

Even if the scope of the pre-made corpus is clearly defined, the inclusion of disparate formats of text can be another artifact compromising the coherence of a corpus. As Kichuk observes, "Although digital repositories continually refer to their text collections, such as [Internet Archive's] Text Archive, as 'book collections,' many of their digitized e-books are not books at all. Many are fragments, pamphlets, even journal articles or book chapters."[37] Reliance on a pre-assembled corpus can also introduce methodological concerns regarding the institutional, cultural, and corpo-

rate biases that have shaped the corpus: Why were these texts preserved and digitized while others were not? The serious CTA researcher may find both predefined corpora and derivative datasets insufficient for research questions that require a clear outline of methods and a detailed understanding of the underlying corpus.

Certainly, there are other corpora that are open for viewing, are relatively easy to assemble, and allow the desired context. Archives from many public memory institutions, such as the Library of Congress's Chronicling America project,[38] provide complete access to full text from their collections. A researcher using optical character recognition (OCR) bulk data downloads from the early American newspapers on the Chronicling America website, for example, can avoid questions about where to access specific newspapers (she will simply use the ones collected by the Library of Congress) and how to legally access them (she will assume the research arm of the US Congress does not offer illegal downloads). Inherently, though, this approach is likely to limit one's sources to older texts that are already in the public domain since these are sometimes the only texts that those institutions are able to legally provide.

## Beyond the Black Box: Fair Use When Building a Corpus from Scratch

As a result of the limitations and artifacts of pre-made corpora, often CTA researchers—like our own scholar of post-WWII poetry—will need to create their own dataset. In doing so, they intrinsically have access to the underlying contents and could read or consume the text if they wanted to. How does building a corpus of copyrighted works from scratch comport with fair use? Would our researcher's use be equally fair if she has access to "consume" the corpus but does not intend to?[39]

Some scholars suggest (as we also do here) that researchers ought to consider the use they are actually making. As more than 100 digital humanities and legal scholars explained in their *amici curiae* brief to the *HathiTrust 2014* court, "Copying to enable purely non-expressive [or non-consumptive] uses, such as the automated extraction of data, does not infringe the statutory rights of the copyright holder."[40] Further, they argued that "if a human's *reading* of copyrighted expression to extract non-expressive material is fair use, the result should be the same when a computer

performs the extraction."[41] Case law supports this construction. Of particular interest is *A.V. ex rel Vanderhye v. iParadigms*, 562 F.3d 630 (4th Cir. 2009) ("iParadigms 2009"), in which defendants made a commercial use of copyrighted works (student papers) to create plagiarism detection software. In affirming summary judgment on defendant software creator's fair use defense, the court held that use of copyrighted works for plagiarism detection had an entirely different function and purpose (i.e., to prevent plagiarism by comparative use) than the expressive content in the original works and was both transformative under factor one and a fair use overall.

Under the *HathiTrust 2014*, *Google Books 2015*, and *iParadigms 2009*, along with other cases that have considered intermediate copying or the creation of searchable databases,[42] building a corpus of poetry used to compute elements of syntax should be found to be equally transformative, even if access to consume is available.[43] Indeed, some cases, like *White v. West Publishing Corp*., 29 F. Supp. 3d 396 (S.D.N.Y. 2014), have even held that preparing a text-searchable, issue coded, and metadata-rich database with copyrighted materials where full access to consume was expressly provided is a transformative use. In *White v. West*, WestLaw and Lexis had included two copyrighted briefs into their text-searchable legal database. Relying on *Campbell v. Acuff-Rose Music*, 510 U.S. 569 (1994), the district court held that "West and Lexis's processes of reviewing, selecting, converting, coding, linking, and identifying the documents 'add…something new, with a further purpose or different character' than the original briefs"[44] and were also a fair use overall even where the database had a commercial purpose.

But will all transformative changes to texts in the creation of corpora for CTA be a fair use overall? Transformativeness under factor one will always still need to be balanced with the three other fair use factors with the understanding that "the more transformative the new work, the less will be the significance of other factors…that may weigh against a finding of fair use."[45]

Moreover, a researcher should understand that while it may be fair use to create and utilize the database for personal research, subsequently *publishing* that database and its substantive contents for others to use (and potentially "consume") may exceed those bounds. Suddenly, one may move from the realm of transformative use (assembly for computational analysis) to pure duplication of copyright-protected texts as the

Second Circuit recently concluded in *Fox News Network, LLC v. TVEyes, Inc.,* 2018 U.S. App. LEXIS 4786 (2nd Cir. Feb. 27, 2018). TVEyes, a media aggregator, records commercial news and radio audiovisual content, imports it into a database, and permits its clients to (among other things) search for, view, download, and share that content in ten-minute clips. Search functionality is made possible because TVEyes copies the closed-captioned text of the content it imports, allowing its clients to search by keyword, date, and time. Fox News Network sued for copyright infringement and, on appeal, the Second Circuit found that enabled features like redistribution exceeded fair use. While keyword-enabled searching would be both transformative and a fair use overall, permitting redistribution was not because it made "available to TVEyes's clients virtually all of Fox's copyrighted content that the clients wish[ed] to see and hear, and because it deprive[d] Fox of revenue."[46]

CTA researchers should thus separately undertake a fair use analysis if they intend to publish excerpts or whole content from the corpora they assemble. If a scholar aims to publish a couple of 500-word annotated excerpts from book-length works so that other scholars may test her algorithm, this may indeed be fair use. Yet, republishing multiple coded chapters for others to work with may not. There is no magic formula here; a researcher's careful consideration of intended uses is key.

## *Contract Law*

Assuming that our CTA researcher has made a fair use of the poems in her creation of a database and is not republishing the database content, has she satisfied law and policy due diligence within the research process?

Indeed, understanding fair use is only one aspect of a scholar's necessary CTA literacies, as contract law can determine what a CTA researcher can do within legal bounds. To illustrate this, suppose our scholar is compiling her digital database of poems from several sources:

- She will download the bulk of the poems from *ProQuest's Literature Online*, a database to which her institution subscribes.
- She will "web scrape" book reviews about the poetry from relevant date ranges within the *New York Times* online.

- For those poems from obscure sources held in print by local archives, she will scan the originals and run optical character recognition (OCR) on them.

She may be surprised to learn that different contracts and agreements govern her access to these materials and what uses she can make of them. A critical early question when attempting to compile a corpus, therefore, is to consider the means by which one has access: is it an institution or library-licensed resource that she is accessing through campus proxies, publicly available on a website, or available via an individual subscription or arrangement with the provider or archives?

## Database License Agreements

Information access is sometimes so seamless that researchers do not realize when they are gaining access to licensed content through library subscriptions.[47] Library subscriptions to licensed content bind authorized users to their terms, even if a researcher was not a party to the library's agreement. To be sure, academic publishers are inclined to enforce these terms: the content is a major commodity for vendors, who charge academic libraries steep (and ever-increasing) sums in exchange for granting access via a license agreement.

Given the potential market value of tightly controlling that content, not all publishers or vendors permit data mining in their license agreements. Standard prohibitory language might be included in a clause labeled "Data Mining" or buried in a paragraph that effectively precludes "downloading all or parts of the content in a systematic or regular manner so as to create a collection of materials."[48] These limitations are possible because, even if building a corpus for research is a fair use, contract law can limit what would otherwise be permitted under federal copyright law.[49]

Academic institutions are beginning to push harder on publishers to permit CTA uses of the licensed materials and may refuse to sign license agreements that, via contract, are end runs around fair use. Successful advocacy may result in text mining clauses that expressly permit bulk downloading, sometimes for "personal research use only," and are often still prohibitive of republishing. As CTA advances, libraries have increasing opportunities (if not obligations) to leverage their role in facilitating

text mining, including by developing better professional advocacy materials to assist other libraries and research institutions.[50]

A significant challenge for researchers in all of this is: How can a researcher discover what her institution's license agreement does or does not permit vis-à-vis text mining? The transparency of such information varies widely across institutions. Library websites and online guides typically offer generalized support regarding CTA but often do not drill down to individual database license terms. The University of British Columbia Library's "License Information" database, however, provides one compelling model for helping users navigate these licenses: the portal allows researchers to search by journal and journal package title and displays information about what CTA (and other) uses are permitted under the university's license agreement covering that resource.[51]

Yet, maintenance of such a database and public-facing portal about license agreements terms requires personnel and resources that not all university libraries are able to provide. Some libraries have instead organized personnel to triage incoming researcher requests. For instance, UC Berkeley's library has a text and data mining e-mail list through which incoming requests reach the library consultants who can advise on various aspects of CTA—including database licensing terms.[52]

In each of these "solutions," there remains a need for literacy education: CTA researchers would need to understand the landscape of library-licensed databases before they knew to ask librarians if their intended use is permitted. Without sufficient outreach on this initial point, it is more likely that systematic or programmatic downloading activity, potentially in violation of library license agreements, will occur. When these violations take place from an IP address for a library proxy, a vendor may block access for all off-campus users by terminating IP access to troubleshoot a potential breach.

## Website Terms of Service

For the poetry reviews our hypothetical researcher is downloading from the *New York Times,* she will not find options to bulk download them from https://nytimes.com. So, she has begun to explore web scraping tools and methods, which have considerable advantages for compiling a text corpus.

Web scraping can automate repetitive tasks such as downloading thousands of PDFs or extracting text from millions of HTML pages via software or code.[53] Unfortunately, scraping articles from https://nytimes.com runs counter to their "terms of service," which appear through a hyperlink in light gray text at the bottom of the *New York Times* website.[54]

Would a court find that our researcher has agreed to be bound by those terms? Some jurisdictions recognize that website "terms of service" or "terms of use" can constitute a valid "browsewrap" agreement.[55] A browsewrap agreement consists of terms and conditions governing use of an internet website that are posted on the website (often accessible by a hyperlink) to which a party assents simply by using the website.[56] This differs from a "clickwrap" agreement, which asks users to check a box affirmatively indicating they assent to the terms provided.[57,58] Browsewrap agreements that contain terms regarding how disputes will be resolved—e.g., mandatory arbitration clauses—or stipulations about attorneys' fees may at times be found unenforceable against public policy in a given jurisdiction. Nevertheless, barring containment of terms that run counter to consumer protections, browsewraps can indeed be valid mechanisms for web content providers to control how their content is used.

Whether a court will enforce the browsewrap, however, depends not only on the jurisdiction and any relevant principles or statutes therein but also the facts of the case. Without a user's actual or constructive knowledge of the terms, courts often do not find the mutual assent required for the formation of a contract.[59] The courts will inquire as to whether the hyperlink to the terms of use is placed in a noticeable location and is of sufficient size, color, font, and more.[60] A few courts have even held that browsewraps cannot be enforceable based solely upon a link to terms at the bottom of a web page.[61]

Given these disparities in browsewrap enforcement, what should a CTA researcher know? A court will look to the facts of whether a researcher had actual or constructive notice, but of course it will do so once a lawsuit has been filed. Needless to say, the mere threat of such a suit can be a deterrent to conducting research. Nor should one advise researchers to bury their heads in the sand and willfully ignore awareness of terms of service—since that fact, too, may be a consideration for the court and constitute constructive awareness.

We encourage researchers to be on notice that terms of service may exist. This is also a best practice because, if a researcher is on campus while crawling a site, it may be a violation of a university or university library's internet policies to violate website terms of use.[62] So, by complying with the website's browsewrap, the researcher also is less likely to run afoul of university policies. Additionally, compliance with both browsewraps and the database license agreements discussed above may eliminate the spectre of potential violations of other statutes like the Computer Fraud & Abuse Act (18 U.S.C. § 1030),[63] the Digital Millennium Copyright Act's anti-circumvention provisions (17 U.S.C. § 1201), and common law rights like trespass to chattel—all of which require further exploration than we are able to undertake here.

A researcher should also consider whether the desired content might actually be available under her institution's licensing agreements, as these agreements can carve out necessary exceptions. For instance, our poetry scholar's library may have negotiated a text and data mining provision in the license agreement that expressly permits CTA. So, if our researcher wanted to crawl the *New York Times* online, this might be disallowed under the public site's terms of use—but permitted via her library's license to ProQuest (a content aggregator).

In addition, and before looking to scrape text from a website, researchers should investigate whether or not the desired data is available via an API or other "framework implemented explicitly to handle and respond to automated data requests."[64] APIs are generally designed to enable commercial reuse that will drive traffic to the content provided via the API host but can also provide a simple, legal point of access for the CTA scholar. In the case of the *New York Times* website, there is an "Article Search API" that might help our scholar identify and access metadata related to reviews of the relevant poetry, but it is not designed to enable full-text access to the *New York Times* archive.[65] Ultimately, therefore, awareness of how a researcher intends to access and download content is another critical step in understanding whether such activity is authorized.

## Agreements with Archives and Special Collections

Our researcher is also digitizing materials from a local archive. All researchers—CTA and otherwise—should be aware that if they are using published or unpublished material from libraries' special collections or archives, they

may need to consider a use agreement they signed with the archives. These agreements typically govern whether the materials can subsequently be published, not whether the scholar can use them in her research *ab initio*.

Why would some libraries and archives restrict one's ability to publish from works in their collections? They may have signed agreements with donors that restrict reuse of the records being contributed. For instance, a donor of unpublished personal letters might—as a condition of donation—restrict use of the letters to researchers in a reading room and prohibit publication or digitization. The archives may pass this condition on to researchers with a "terms of use" (or equivalent) agreement. In exchange for the archives granting access to the correspondence, researchers may waive rights to publish excerpts from the letters, even if doing so would be fair use. Here, too, we also face the distinction between permission to use for research and permission to republish in digital scholarship.

The good news is that archives and cultural heritage institutions are often willing to engage in discussion about expanding permissible uses; at a minimum, it is worth asking the institution if the intended use can be allowed. Indeed, sometimes the archives' "terms of use" agreement is more restrictive than even the archives had intended while institutional practice catches up to the growing landscape for how to describe reuse rights.[66]

## *Ethics*

Lastly, even in cases where scraping text from a site may be permissible under the Copyright Act, a database license agreement, and a website's terms of use, a researcher should also consider best practices regarding the impact of programmatically downloading or indexing content from a web server. As Munzert et al. detail, "Maintainers of websites sometimes want to keep at least some of their content prohibited from being crawled, for example, to keep their server traffic in check. This is what the robots.txt file is used for. This 'Robots Exclusion Protocol' tells the robots which information on the site may be harvested."[67]

Robots.txt files and the "robots" <meta> tag in HTML headers are designed primarily to tell search engines when web crawlers used to index a site for public retrieval are prohibited or allowed. While prohibitions

in robots.txt fall into a legal gray area and may not explicitly forbid web scraping, it is a generally accepted best practice that any programmatic access to a site respects the wishes of the web host.[68]

# Toward a Workflow

How can we transform these principles into a workflow adaptable and adoptable by CTA researchers in the field who are using or building corpora? We believe it may be helpful to interweave these literacies into three stages of outreach and education: use of precompiled corpora, corpus creation, and corpus publishing.

## *Use of Precompiled Corpora*

Many libraries provide online text mining guides listing open access and licensed resources that are available to their users for CTA purposes.[69] In this section of the workflow, however, rather than focus on where researchers should *look* for these corpora, we enumerate the literacy considerations they should make when choosing to *use* any pre-compiled corpora.

### Address Scope of the Corpus

A researcher should be able to articulate the boundaries of her corpus, not simply in terms of the total number of items represented, but taking into consideration the original sources represented in the collection and the granularity with which one can query subsets of the collection. While black box corpora may not allow for traditional consumption of texts, platforms offer various levels of access to information about the objects in their collections. Consideration should also be given to what may not be included in the corpus and whether or not those items were left out for reasons related to copyright, privacy, or other legal policy.

### Consider Legal Frameworks Shaping Corpus Format or Contents

Relatedly, if a researcher uses a corpus shaped by legal contours like copyright or privacy statutes and agreements, she should understand how

these factors can limit potential uses of the underlying texts. For instance, (1) familiarity with fair use law illustrates why Google makes only Snippets views of certain works available or why HDL allows full-text searching but not full-content viewing; (2) similarly, an understanding of privacy law can help a researcher explain why various items were not viewable or were excluded entirely from a given corpus.

#### Account for Mode of Access

While CTA researchers may initially seek out access to bulk downloads of familiar representations of texts from any given collection, familiarity with emerging modes of access enabled by fair use—including derived downloadable datasets, secure computing environments, and web-based tools for interacting with a corpus—will open up significant new opportunities for access. It is equally important for researchers to recognize which modes of access allow specific research methods and questions that they are hoping to pursue.

#### Explore Digital File History and Metadata

The CTA researcher should consider how and when a particular corpus has been digitized, the degree to which it represents the original objects in the collection—considering, for example, the quality of OCR—and the role that various institutions played in defining and digitizing its contents. Along the same lines, researchers should pay careful attention to the quality and kinds of metadata provided for individual items in a corpus. Without accurate bibliographic publication dates, for example, a researcher will be unable to perform temporal analyses of items in the corpus.[70]

### *Corpus Creation*

A CTA researcher who seeks to develop a corpus must rely on additional literacies that integrate a more nuanced understanding of copyright and licensing that librarians are well-suited to provide.

#### Consider Copyright and Fair Use Rights

Researchers should be equipped to consider whether the content of the corpus they build is protected by copyright and, if so, whether it would

be fair use to create a searchable database of these materials under *HathiTrust 2014*, *Google Books 2015*, and other cases. Creating a research database for personal use or non-consumptive text mining has typically been found to be fair, though the fair use balancing test may yield a different outcome with respect to publishing from those corpora (as noted below).

### Assess Means of Content Access

- Via institutional license agreement. Sufficient outreach to researchers should occur such that they have an understanding of when they are actually utilizing library- or institution-licensed resources and databases. With that understanding comes awareness of the need to discern (1) whether the license curbs uses that would otherwise qualify as fair use, and (2) whether that license permits text and data mining and the creation of a collection. Scholars should also understand that web scraping in violation of a database license agreement might, if done on campus, also impedes access to the database for other campus users. Here, again, librarian contact or information provision is key.

- Via website. Before compiling a corpus via a website, a researcher should consider whether the same content is available through her institution's licensed databases, as the license agreements may expressly allow CTA even if the "vanilla" usage terms of a website bearing that same content do not. If, however, a researcher is indeed using materials on the open web and not through an institutionally licensed resource, useful literacies include:

    - understanding the scope of and permissible uses defined by a website's "terms of service" or "terms of use," or any other "browsewrap" or "clickwrap" licenses they may be deemed to have entered into by using the site for their intended purpose;
    - understanding of formal web services for legal access to web content—before researchers consider building or using a web scraper, they should investigate whether or not the platform of interest offers an Application Programming Interface (API) or other programmatic or bulk access point to the content they need; and

    - consideration of best practices concerning programmatic access, including the limitations and prohibitions documented in a site's robots.txt and "robots" <meta> tag—researchers should be cognizant that large download requests can impact server performance negatively and bear financial costs for content providers.

- Via archives, museum, or library special collection. All researchers, CTA, and consumptive readers alike should be aware of any use or republishing restrictions they may be asked to accept when acquiring copies of materials from library special collections, archives, or museums. This is separate from any underlying copyright attached to the materials themselves. Before signing any agreements with memory institutions, researchers should therefore consider what uses they intend to make of the content and be prepared to ask the institution—in writing—for permission to store or publish the content in ways that satisfy those intended uses. Researchers should keep records of any permissions obtained.

## *Corpus Publishing*

CTA researchers' end goals may be not only to publish scholarship with their findings but also publish the raw, annotated, or coded content itself. Publishing the content or the database they have created helps other scholars test their own algorithms and provides raw material upon which to conduct their own research and tests. Yet, it is often in the republishing of the corpus content that the limits of fair use are reached or the bounds of license agreements are exceeded. CTA researchers should thus separately undertake a fair use analysis if they intend to publish excerpts or whole content from the corpora they assemble.

Once again, the same license agreements and browsewraps can infuse additional parameters for what may be included as republished content. Typically, academic libraries negotiate agreements that allow for quoting or excerpting materials within the bounds of fair use—so, potentially, the researcher's intended republication may very well fall within the contours

of what the license agreement allows. Visibility into what agreements an institution has signed once again remains important.

## From Workflows to Skill Sets

Reflecting upon our digital humanities scholar studying post-WWII poetry, in the light of the literacies unmasked by our workflow, we recognize this as a call to action. Working at the intersection of copyright law, database licensing, and public service, academic libraries are increasingly well-equipped to support CTA scholars throughout the research lifecycle. We encourage coordinated efforts by professional library organizations in the US to help institutions operationalize literacy workflows, such as the one outlined above, so that CTA scholars may build requisite skill sets to support their research.

As we have indicated throughout, core literacies would help CTA researchers to recognize, among other things, that copyright fair use jurisprudence affects how a corpus might appear, whether a researcher can create a corpus from scratch, and whether she may subsequently share it; contract law (including agreements that researchers may not have personally signed or agreed to) can supplant these fair use rights; community ethics may influence best practices for content aggregation; and, finally, that there may be other considerations for CTA researchers in the use, creation, and publishing of corpora—such as questions of privacy and publicity rights or matters invoking indigenous knowledge that are beyond the scope of what we cover here.

Perhaps the key literacy, therefore, is for CTA researchers to understand the need for a workflow itself and to explore a tailored approach in consultation with their librarians. Achieving this fundamental literacy necessitates outreach and education on the issues identified here to bring a scholar to the stage at which she could apply statistical computing methods on a robust and lawfully assembled corpus.

# Endnotes

1. The authors would like to thank the following scholars for their careful review and valuable feedback on earlier drafts of this chapter: Eleanor Dickson, HTRC Digital Humanities Specialist at the University of Illinois at Urbana-Champaign; Michael Wolfe, Scholarly Communications Officer, UC Davis Library; and David J. Hansen, Director of Copyright and Scholarly Communications, Duke University Library.
2. Mathew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Champaign, IL: University of Illinois Press, 2013), 15.
3. Marti A. Hearst, "What is Text Mining?," last modified October 17, 2003, http://people.ischool.berkeley.edu/~hearst/text-mining.html; Martin Truyens and Patrick Van Eecke, "Legal Aspects of Text Mining," *Computer Law & Security Review* 30, no. 2 (April 2014): 153.
4. Truyens and Van Eecke, "Legal Aspects of Text Mining"; Jonathan Clark, *Text Mining & Scholarly Publishing* (Loosdrecht, Netherlands: Publishing Research Consortium, 2012), https://www.stm-assoc.org/2012_01_01_PRC_Clark_Text_Mining_and_Scholarly_Publishing.pdf; Intellectual Property Office, *Exceptions to Copyright: Research*, last modified November 18, 2014, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.
5. Michael L. Black, "The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond Through Internet Research," *International Journal of Humanities and Arts Computing* 10, no. 1 (March 2016), https://doi.org/10.3366/ijhac.2016.0162; Laura Quilter, "Copyright Futures in the Digital Humanities," presentation at the UMass Digital Humanities Initiative, Amherst, MA, March 11, 2013, https://works.bepress.com/laura_quilter/27/; Ryan Mitchell, *Web Scraping With Python: Collecting Data from the Modern Web* (Sebastopol, CA: O'Reilly Media, 2015), chap. 14, pt. C, "The Legalities and Ethics of Web Scraping."
6. U.S. Const. Art. 1, § 8.; see also Authors Guild v. Google, 804 F.3d 202, 209 (2nd Cir. 2015).
7. 804 F.3d at 209.
8. Copyright Act of 1976, 17 U.S.C. § 107.
9. Vanderhye v. iParadigms, LLC, 562 F.3d 630, 640 (4th Cir. 2009) (quoting Stewart v. Abend, 495 U.S. 2017, 237 (1990)).
10. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 578 (1994).
11. *Campbell* at 591; 562 F.3d at 643.
12. 510 U.S. at 578–79.
13. 562 F.3d at 639.
14. James Grimmelman, "Copyright for Literate Robots," *Iowa Law Review* 101 (2016): 659; Pierre N. Leval, "Toward a Fair Use Standard," *Harvard Law Review* 103, no. 5 (March 1990): 1111, https://doi.org/10.2307/1341457.
15. 510 U.S. at 579.
16. 562 F.3d 630 (quoting Bond v. Blum, 317 F.3d 385, 396 (4th Cir. 2003)).
17. 562 F.3d at 643 (quoting NXIVM Corp. v. The Ross Institute, 364 F.3d 471, 482 (2nd Cir. 2004)).
18. 510 U.S. at 592; 562 F.3d at 643.
19. 804 F.3d at 209.
20. For additional review of various cases that treat copyright and text mining, see Krista

L. Cox, "Research Libraries and New Technologies, Promoting Access to Information, Learning, and Innovation for Today and the Future," *I/S: Journal of Law and Policy for the Information Society* 13, no. 1 (2016), http://hdl.handle.net/1811/81137.

21. HathiTrust "is a partnership of major research institutions" committed to "contributing to research, scholarship, and the common good by collaboratively collecting, organizing, preserving, communicating, and sharing the record of human knowledge." HathiTrust's Digital Library "is a digital preservation repository and highly functional access platform. It provides long-term preservation and access services for public domain and in copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives." See https://www.hathitrust.org/about.
22. HathiTrust has since released additional options, like a secure computing environment, through the HathiTrust Research Center tools (discussed above), that allow even more robust CTA to be performed—including visualizations, tables, quotations, or other transformations from the copyrighted materials in the archive.
23. 755 F.3d at 91.
24. Ibid. at 96.
25. Ibid. at 97.
26. Ibid. at 100.
27. Ibid. at 98–99.
28. *Cert. den'd* (No. 15-849) 136 S. Ct. 1658 (Apr. 18, 2016).
29. Matthew L. Jockers, Matthew Sag, and Jason Schultz, "Don't Let Copyright Block Data Mining," *Nature* 490 (October 2012): 30, https://doi.org/10.1038/490029a.
30. "HTRC Data Capsule," HathiTrust Research Center Documentation, accessed on October 31, 2017, https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule.
31. "Data for Research," JSTOR, accessed September 10, 2017, https://dfr.jstor.org/.
32. Boris Capitanu et al, "The HathiTrust Research Center Extracted Feature Dataset (1.0)" (dataset), HathiTrust Research Center, http://dx.doi.org/10.13012/J8X63JT3.
33. "Ngram Viewer," Google Books, accessed on September 10, 2017, http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.
34. Alexander Koplenig, "The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets—Reconstructing the Composition of the German Corpus in Times of WWII," *Digital Scholarship in the Humanities* 32, no. 1 (April 2017), https://doi.org/10.1093/llc/fqv037.
35. For a discussion of the ways in which the lack of bibliographic metadata in Google Books renders the search tool unsuitable for many scholarly purposes, see Geoffrey Nunberg, "Google's Book Search: A Disaster for Scholars," *Chronicle of Higher Education*, August 31, 2009, http://www.chronicle.com/article/Googles-Book-Search-A/48245/.
36. David M. Blei, "Probabilistic Topic Models," *Communications of the ACM* 55, no. 4 (April 2012): 82, https://doi.org/10.1145/2133806.2133826.
37. Diana Kichuk, "Loose, Falling Characters and Sentences: The Persistence of the OCR Problem in Digital Repository E-Books," *portal: Libraries and the Academy* 15, no. 1 (January 2015): 66–67, https://doi.org/10.1353/pla.2015.0005.
38. "Chronicling America," Library of Congress, accessed October 13, 2017, http://chroniclingamerica.loc.gov/.
39. This question was on the radar, too, for the Association of Research Libraries in its *Code of Best Practices in Fair Use for Academic & Research Libraries*, which was

released in 2014, prior to HathiTrust 2014 and Google Books 2015. The *Code of Best Practices* urges libraries creating copyright-protected corpora to limit access to researchers only under non-consumptive terms. (Washington DC: Association of Research Libraries, 2012), http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf.

40. Brief of Digital Humanities and Law Scholars as *Amici Curiae*, Authors Guild v. HathiTrust, No. 12-4547-cv (2d Cir., June 4, 2013), 24.
41. Brief of Digital Humanities and Law Scholars, at 27.
42. See also Kelly v. Arriba Soft, 336 F.3d 811 (9th Cir. 2003) (fair use for search engine to include thumbnails and in-line linking to images hosted on photographer website); Perfect 10 v. Amazon, 508 F. 3d. 1146 (9th Cir. 2007) (Google search engine's use of copyrighted images in the form of thumbnails and in-line linking to full images is transformative fair use); Field v. Google, 412 F. Supp. 2d 1106 (D. Nev. 2006) (Google's cached copies of copyrighted website content used for web page or archival content comparisons, or identification of search query terms, is fair use).
43. The judge who authored *Google Books 2015* would seem to agree with this determination. In 1990, Pierre Leval wrote: "If [a] secondary use adds value to the original—if the quoted matter is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings—this is the very type of activity that the fair use doctrine intends to protect for the enrichment of society." "Toward a Fair Use Standard," 1111.
44. 29 F.Supp.3d at 399.
45. 510 U.S. at 579; HathiTrust 2014.
46. 2018 U.S. App. LEXIS 4786, at *8.
47. Jeremy Frumkin and Terry Reese, "Provision Recognition: Increasing Awareness of the Library's Value in Delivering Electronic Information Resources," *Journal of Library Administration* 51, no. 7–8 (October 2011): 811, http://dx.doi.org/10.1080/01930826.2011.601277.
48. See, e.g., "Terms and Conditions," ProQuest, accessed on October 11, 2017, http://www.proquest.com/about/terms-and-conditions.html; "EBSCO License Agreement," EBSCO, accessed on October 11, 2017, https://www.ebsco.com/terms-of-use.
49. See, e.g., Bowers v. Baystate Techs., 320 F.3d 1317, 1325–1326 (Fed. Cir. 2003); Melville B. Nimmer and David Nimmer, *Nimmer on Copyright*, § 1.01[B][2] (New York: Matthew Bender, 2018).
50. See, e.g., Leslie A. Williams et al., "Negotiating a Text Mining License for Faculty Researchers," *Information Technology and Libraries* 33, no. 3 (September 2014).
51. "License Information," University of British Columbia, Library, accessed on October 12, 2017, https://licenses.library.ubc.ca/.
52. Emailing tdm-access@berkeley.edu reaches all library staff involved in supporting CTA research, including, for instance, the e-resources librarian, e-learning librarian, scholarly communication officer, digital humanities librarian, and more.
53. Further, open-source code to scrape websites is easy to find. As of this writing, there are over 24,000 results for a search for "scraper" on the code repository, GitHub.com, for example. Popular guides often explain technical methods to collect and organize text from websites. See e.g., Mitchell, *Web Scraping with Python*; Simon Munzert et al., *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (Chichester, West Sussex, UK: Wiley, 2014), Wiley Online Library.
54. "Terms of Service," *New York Times*, accessed on October 17, 2017, https://www.nytimes.com/content/help/rights/terms/terms-of-service.html.

55. See, e.g., E.K.D. ex rel. Dawes v. Facebook, 885 F.Supp. 2d 894 (S.D. IL. 2012) (applying California law); Snap-on Business Solutions v. O'Neil & Assoc., 708 F.Supp. 2d 669 (N.D. Ohio 2010) (apparently applying Ohio law); see also Allison S. Brehm and Cathy D. Lee, "Click Here to Accept the Terms of Service," *Communications Lawyer* (American Bar Association) 31, no. 1 (January 2015), https://www.americanbar.org/publications/communications_lawyer/2015/january/click_here.html.
56. Be In v. Google, 2013 WL 5568706, at *6 (N.D. Cal. Oct. 9, 2013); "Validity, Construction, and Application of Browsewrap Agreements," 95 A.L.R. 6th 57.
57. 95 A.L.R. 6th 57.
58. If a researcher engages in her own subscription to a particular web resource, she may very well be asked to enter into a clickwrap agreement, expressly assenting to be bound by the web provider's terms.
59. See, e.g., 2013 WL 5568706, at *7 (N.D. Cal., 2013) ("Most courts upholding the enforceability of browsewrap agreements have done so in circumstances where notice to the defendant was firmly established in the factual record."); Specht v. Netscape Communications, 306 F.3d 17 (2nd Cir. 2002) (applying CA and NY law) ("a reasonably prudent offeree in plaintiffs' position would not have known or learned, prior to acting on the invitation to download, of the reference to [the] license terms hidden below the 'Download' button on the next screen").
60. See, e.g., Long v. Provide Commerce, 200 Cal. Rptr. 3d 117, 125–26, (Cal. App. 2016) ("Here, the Terms of Use hyperlinks—their placement, color, size and other qualities relative to the ProFlowers.com Web site's overall design—are simply too inconspicuous to" warrant assent to an agreement.); Cf. Cairo v. Crossmedia Services, No. 04–04825, 2005 WL 756610 (N.D. Cal., Apr. 1, 2005) (every page on the website at issue had a text notice that read: "By continuing past this page and/or using this site, you agree to abide by the Terms of Use for this site, which prohibit commercial use of any information on this site.").
61. 2013 WL 5568706, at *7 (N.D. Cal., 2013) ("At least one court has found that actions seeking to enforce website terms of use as an enforceable browsewrap contract must allege more than the mere existence of a link at the bottom of a page." [citing Cvent, Inc. v. Eventbrite, 739 F.Supp. 2d 927, 936 (E.D. Va. 2010); Nguyen v. Barnes & Noble, 12–CV–0812, 2012 WL 3711081 (C.D. Cal. Aug. 28, 2012) (refusing to enforce arbitration agreement where notice of browsewrap agreement was predicated merely on a link at the bottom of the website)]).
62. See, e.g., the "Conditions of Use and Licensing Restrictions for Electronic Resources," University of California, Berkeley, Library, accessed October 17, 2017, http://www.lib.berkeley.edu/about/conditions-of-use-for-electronic-resources.
63. For an overview of how courts have applied the Computer Fraud & Abuse Act, see Buckman, Annotation, "Validity, Construction, and Application of Computer Fraud and Abuse Act (18 U.S.C.A. § 1030)."
64. Black, "The World Wide Web," 98.
65. "The New York Times Developers," *New York Times*, accessed on October 16, 2017, https://developer.nytimes.com/.
66. Rightsstatements.org is one project aimed at expanding ways in which cultural heritage institutions can describe both copyright status and reuse rights for their digital items and materials—thus removing often unintended binary language constricting researchers' usage rights. For a discussion of alternative approaches, institutions concerned about whether their archives agreements are overreaching should also see Kenneth D. Crews,

"Museum Policies and Art Images: Conflicting Objectives and Copyright Overreaching," *Fordham Intellectual Property, Media & Entertainment Law Journal* 22 (July 1, 2012), https://ssrn.com/abstract=2120210.

67. Munzert et al., *Automated Data Collection with R*, 280.
68. Black, "The World Wide Web."
69. See, e.g., "Text Mining & Computational Analysis," University of California, Berkeley, Library, accessed on September 10, 2017, https://guides.lib.berkeley.edu/text-mining; "Text and Data Mining," University of Chicago, Library, accessed on September 10, 2017, http://guides.lib.uchicago.edu/textmining.
70. For a closer consideration of challenges related to publication dates as represented in bibliographic metadata for large-scale corpora analysis, see David Bamman et al., "Estimating the Date of First Publication in a Large-Scale Digital Library," in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries, Toronto, Canada, June 2017 (JCDL'17)* (Piscataway, NJ: IEEE, 2017), https://doi.org/10.1109/JCDL.2017.7991569.

# Bibliography


Association of Research Libraries. *Code of Best Practices in Fair Use for Academic and Research Libraries*. Washington DC: Association of Research Libraries, 2012. http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf.

Bamman, David, Michelle Carney, Jon Gillick, and Cody Hennesy. "Estimating the Date of First Publication in a Large-Scale Digital Library." In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries, Toronto, Canada, June 2017 (JCDL'17)*. Piscataway, NJ: IEEE, 2017. https://doi.org/10.1109/JCDL.2017.7991569.

Black, Michael L. "The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond Through Internet Research." *International Journal of Humanities and Arts Computing* 10, no. 1 (March 2016): 95–109. https://doi.org/10.3366/ijhac.2016.0162.

Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4 (April 2012): 77–84. https://doi.org/10.1145/2133806.2133826.

Brehm, Allison S., and Cathy D. Lee. "Click Here to Accept the Terms of Service." *Communications Lawyer* (American Bar Association) 31, no. 1 (January 2015): 4–7. https://www.americanbar.org/publications/communications_lawyer/2015/january/click_here.html.

Brief of Digital Humanities and Law Scholars as *Amici Curiae*, Authors Guild v. HathiTrust, No. 12-4547-cv (2d Cir., June 4, 2013).

Buckman, Deborah F. Annotation, *Validity, Construction, and Application of Computer Fraud and Abuse Act (18 U.S.C.A. § 1030)*, 174 A.L.R. Fed. 101 (2017).

Capitanu, Boris, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, and J. Stephen Downie. "The HathiTrust Research Center Extracted Feature Dataset (1.0)" (dataset). *HathiTrust Research Center*. http://dx.doi.org/10.13012/J8X63JT3.

Clark, Jonathan. *Text Mining & Scholarly Publishing.* Loosdrecht, Netherlands: Publishing Research Consortium, 2012. https://www.stm-assoc.org/2012_01_01_PRC_Clark_Text_Mining_and_Scholarly_Publishing.pdf.

Copyright Act of 1976, 17 U.S.C. § 107 (2012).

Cox, Krista L. "Research Libraries and New Technologies, Promoting Access to Information, Learning, and Innovation for Today and the Future." *I/S: Journal of Law*

*and Policy for the Information Society* 13, no. 1 (2016): 261–94. http://hdl.handle.net/1811/81137.

Crews, Kenneth D. "Museum Policies and Art Images: Conflicting Objectives and Copyright Overreaching." *Fordham Intellectual Property, Media & Entertainment Law Journal* 22 (July 1, 2012): 795–834. https://ssrn.com/abstract=2120210.

EBSCO. "EBSCO License Agreement." Accessed on October 11, 2017. https://www.ebsco.com/terms-of-use.

E.K.D. ex rel. Dawes v. Facebook, Inc., 885 F. Supp. 2d 894 (S.D. Ill. 2012).

Frumkin, Jeremy, and Terry Reese. "Provision Recognition: Increasing Awareness of the Library's Value in Delivering Electronic Information Resources." *Journal of Library Administration*, 51, no. 7–8 (October 2011): 810–19. http://dx.doi.org/10.1080/01930826.2011.601277.

Google Books. "Ngram Viewer." Accessed on September 10, 2017. http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

Grimmelman, James. "Copyright for Literate Robots." *Iowa Law Review* 101 (2016): 657–81. https://ilr.law.uiowa.edu/print/volume-101-issue-2/copyright-for-literate-robots/.

Harper, Blaney, and Vaishali Udupa. "Drafting Electronic Software Licenses to Prevent Reverse Engineering." *E-Commerce Law Report* 6, no. 2 (January 2004).

HathiTrust. "About." Accessed on March 13, 2018. https://www.hathitrust.org/about.

Hearst, Marti A. "Untangling Text Data Mining." In *ACL '99 Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics,* 3–10. College Park: Association for Computational Linguistics, 2010. https://doi.org/10.3115/1034678.1034679.

———. "What is Text Mining?" Last modified October 17, 2003. http://people.ischool.berkeley.edu/~hearst/text-mining.html.

"HTRC Data Capsule." *HathiTrust Research Center Documentation.* Accessed on October 31, 2017. https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule.

Intellectual Property Office. *Exceptions to Copyright: Research.* Last modified November 18, 2014. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.

Jockers, Matthew L., Matthew Sag, and Jason Schultz. "Don't Let Copyright Block Data Mining." *Nature* 490 (October 2012): 29–30. https://doi.org/10.1038/490029a.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. Champaign, IL: University of Illinois Press, 2013. https://muse.jhu.edu/book/21978.

JSTOR. "Data for Research." Accessed September 10, 2017. https://dfr.jstor.org/.

Kemper, Kurtis A. *Annotation, Validity, Construction, and Application of Browsewrap Agreements*, 95 A.L.R. 6th 57 (2017).

Kichuk, Diana. "Loose, Falling Characters and Sentences: The Persistence of the OCR Problem in Digital Repository E-Books." *portal: Libraries and the Academy* 15, no. 1 (January 2015): 59–91. https://doi.org/10.1353/pla.2015.0005.

Koplenig, Alexander. "The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets—Reconstructing the Composition of the German Corpus in Times of WWII." *Digital Scholarship in the Humanities* 32, no. 1 (April 2017): 169–88. https://doi.org/10.1093/llc/fqv037.

Leval, Pierre N. "Toward a Fair Use Standard." *Harvard Law Review* 103, no. 5 (March 1990): 1105–36. https://doi.org/10.2307/1341457.

Library of Congress. "Chronicling America." Accessed October 13, 2017. http://chroniclingamerica.loc.gov/.

Mitchell, Ryan. *Web Scraping with Python: Collecting Data from the Modern Web*. Sebastopol, CA: O'Reilly Media, 2015. Safari Books Online.

Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Chichester, West Sussex, UK: Wiley, 2014. Wiley Online Library.

*New York Times*. Accessed on October 12, 2017. https://www.nytimes.com/.

———. "The New York Times Developers." Accessed on October 16, 2017. https://developer.nytimes.com/.

———."Terms of Service." Accessed on October 17, 2017. https://www.nytimes.com/content/help/rights/terms/terms-of-service.html.

Nimmer, Melville B., and David Nimmer. *Nimmer on Copyright* § 1.01[B][2]. New York: Matthew Bender, Rev. Ed, 2018.

Nunberg, Geoffrey. "Google's Book Search: A Disaster for Scholars." *Chronicle of Higher Education* (August 31, 2009). http://www.chronicle.com/article/Googles-Book-Search-A/48245/.

ProQuest. "Terms and Conditions." Accessed on October 11, 2017. http://www.proquest.com/about/terms-and-conditions.html.

Quilter, Laura. "Copyright Futures in the Digital Humanities." Presentation at the UMass Digital Humanities Initiative, Amherst, MA, March 11, 2013. https://works.bepress.com/laura_quilter/27/.

Truyens, Martin, and Patrick Van Eecke. "Legal Aspects of Text Mining." *Computer Law & Security Review* 30, no. 2 (April 2014): 153–70. https://doi.org/10.1016/j.clsr.2014.01.009.

University of British Columbia, Library. "License Information." Accessed on October 12, 2017. https://licenses.library.ubc.ca/.

University of California, Berkeley, Library. "Conditions of Use and Licensing Restrictions for Electronic Resources." Accessed October 17, 2017. http://www.lib.berkeley.edu/about/conditions-of-use-for-electronic-resources.

———. "Text Mining & Computational Analysis." Accessed on September 10, 2017. http://guides.lib.berkeley.edu/text-mining.

University of Chicago, Library. "Text and Data Mining." Accessed on September 10, 2017. http://guides.lib.uchicago.edu/textmining.

Williams, Leslie A., Lynne M. Fox, Christophe Roeder, and Lawrence Hunter. "Negotiating a Text Mining License for Faculty Researchers." *Information Technology and Libraries* 33, no. 3 (September 2014): 5–21.